SS 2004

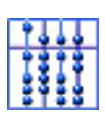
Diskrete Strukturen II

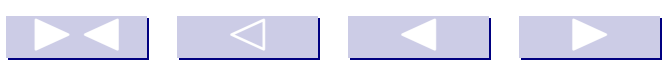
Ernst W. Mayr

Fakultät für Informatik

TU München

http://www14.in.tum.de/lehre/2004SS/ds/index.html.de





Weiterer Beweis von Satz 44:

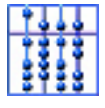
Beweis: Gemäß dem vorhergehenden Beispiel gilt

$$M_{X_i}(t) = e^{t\mu_i + (t\sigma_i)^2/2}\,.$$

Wegen der Unabhängigkeit der X_i folgt

$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{t(a_1X_1+\cdots+a_nX_n)}] = \prod_{i=1}^{n} \mathbb{E}[e^{(a_it)X_i}] \\ &= \prod_{i=1}^{n} M_{X_i}(a_it) \\ &= \prod_{i=1}^{n} e^{a_it\mu_i + (a_it\sigma_i)^2/2} \\ &= e^{t\mu + (t\sigma)^2/2}, \end{aligned}$$

mit $\mu = a_1\mu_1 + \cdots + a_n\mu_n$ und $\sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$. *q. e. d.*



2.4 Zentraler Grenzwertsatz

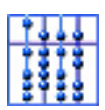
Satz 45: (Zentraler Grenzwertsatz)

Die Zufallsvariablen $X_1, \ldots, X_n$ besitzen jeweils dieselbe Verteilung und seien unabhängig. Erwartungswert und Varianz von X_i existieren für $i = 1, \ldots, n$ und seien mit μ bzw. σ^2 bezeichnet ($\sigma^2 > 0$).

Die Zufallsvariablen Y_n seien definiert durch $Y_n := X_1 + \ldots + X_n$ für $n \geq 1$. Dann folgt, dass die Zufallsvariablen

$$Z_n := \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

asymptotisch standardnormalverteilt sind, also $Z_n \sim \mathcal{N}(0,1)$ für $n \longrightarrow \infty$.

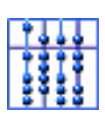


Etwas formaler ausgedrückt gilt: Die Folge der zu Z_n gehörenden Verteilungsfunktionen F_n hat die Eigenschaft

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \text{ für alle } x \in \mathbb{R}.$$

Wir sagen dazu auch: Die Verteilung von Z_n konvergiert gegen die Standardnormalverteilung für $n \to \infty$.

Dieser Satz ist von großer Bedeutung für die Anwendung der Normalverteilung in der Statistik. Der Satz besagt, dass sich die Verteilung einer Summe beliebiger unabhängiger Zufallsvariablen (mit endlichem Erwartungswert und Varianz) der Normalverteilung umso mehr annähert, je mehr Zufallsvariablen an der Summe beteiligt sind.



Beweis:

Wir betrachten $X_i^* := (X_i - \mu)/\sigma$ für $i = 1, \ldots, n$ mit $\mathbb{E}[X_i^*] = 0$ und $\mathrm{Var}[X_i^*] = 1$. Damit gilt (gemäß vorhergehendem Beispiel)

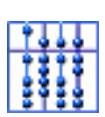
$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t(X_1^* + \ldots + X_n^*)/\sqrt{n}}] \\ &= M_{X_1^*}(t/\sqrt{n}) \cdot \ldots \cdot M_{X_n^*}(t/\sqrt{n})\,. \end{aligned}$$

Für beliebiges i betrachten wir die Taylorentwicklung von $M_{X_i^*}(t) =: h(t)$ an der Stelle $t = 0$

$$h(t) = h(0) + h'(0) \cdot t + \frac{h''(0)}{2} \cdot t^2 + \mathcal{O}(t^3).$$

Aus der Linearität des Erwartungswerts folgt

$$h'(t) = \mathbb{E}[e^{tX_i^*} \cdot X_i^*] \text{ und } h''(t) = \mathbb{E}[e^{tX_i^*} \cdot (X_i^*)^2].$$



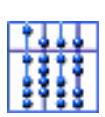
Damit gilt

$$h'(0) = \mathbb{E}[X_i^*] = 0 \text{ und } h''(0) = \mathbb{E}[(X_i^*)^2] = \mathrm{Var}[X] = 1.$$

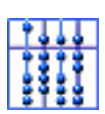
Durch Einsetzen in die Taylorreihe folgt $h(t) = 1 + t^2/2 + \mathcal{O}(t^3)$, und wir können $M_Z(t)$ umschreiben zu

$$M_Z(t) = \left(1 + \frac{t^2}{2n} + \mathcal{O}\left(\frac{t^3}{n^{3/2}}\right)\right)^n \to e^{t^2/2} \text{ für } n \to \infty.$$

Aus der Konvergenz der momenterzeugenden Funktion folgt auch die Konvergenz der Verteilung. Damit ist Z asymptotisch normalverteilt.

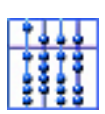


Die momenterzeugende Funktion existiert leider nicht bei allen Zufallsvariablen und unser Beweis ist deshalb unvollständig. Man umgeht dieses Problem, indem man statt der momenterzeugenden Funktion die so genannte charakteristische Funktion $\tilde{M}_X(t) = \mathbb{E}[e^{\mathrm{i}tX}]$ betrachtet. Für Details verweisen wir auf die einschlägige Literatur. *q. e. d.*



Der Zentrale Grenzwertsatz hat die folgende intuitive Konsequenz:

> Wenn eine Zufallsgröße durch lineare Kombination vieler unabhängiger, identisch verteilter Zufallsgrößen entsteht, so erhält man näherungsweise eine Normalverteilung.



Ein wichtiger Spezialfall das Zentralen Grenzwertsatzes besteht darin, dass die auftretenden Zufallsgrößen Bernoulli-verteilt sind.

Korollar 46: (Grenzwertsatz von DeMoivre)

$X_1, \ldots, X_n$ seien unabhängige Bernoulli-verteilte Zufallsvariablen mit gleicher Erfolgswahrscheinlichkeit p. Dann gilt für die Zufallsvariable H_n mit

$$H_n := X_1 + \ldots + X_n$$

für $n \geq 1$, dass die Verteilung der Zufallsvariablen

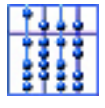
$$H_n^* := \frac{H_n - np}{\sqrt{np(1-p)}}$$

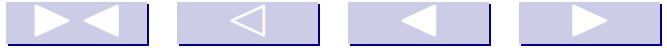
$n \longrightarrow \infty$ gegen die Standardnormalverteilung konvergiert.

Beweis: Die Behauptung folgt unmittelbar aus dem Zentralen Grenzwertsatz, da $\mu = \mathbb{E}[H_i] = p$ und $\sigma^2 = \mathrm{Var}[H_i] = p(1-p)$.
q. e. d.

Bemerkung

Wenn man $X_1, \ldots, X_n$ als Indikatorvariablen für das Eintreten eines Ereignisses A bei n unabhängigen Wiederholungen eines Experimentes interpretiert, dann gibt H_n die absolute Häufigkeit von A an.

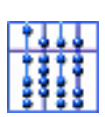


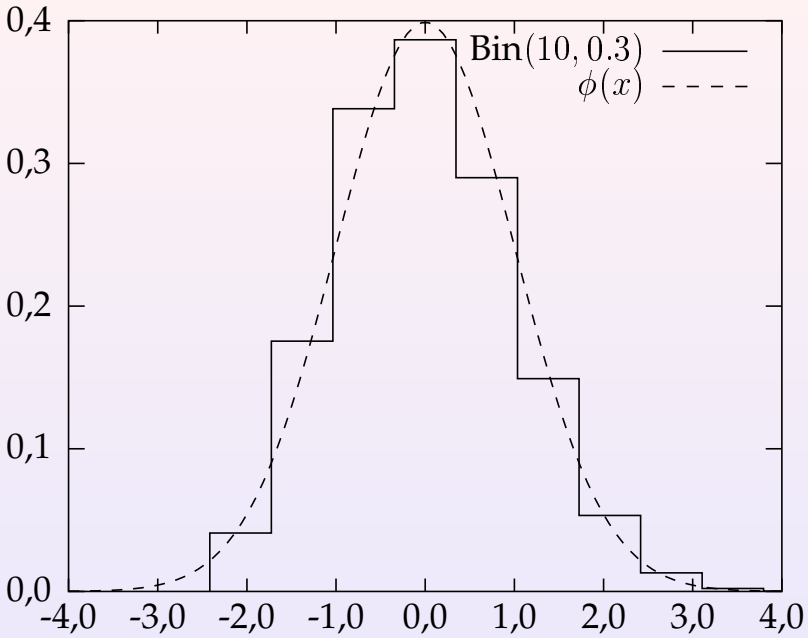


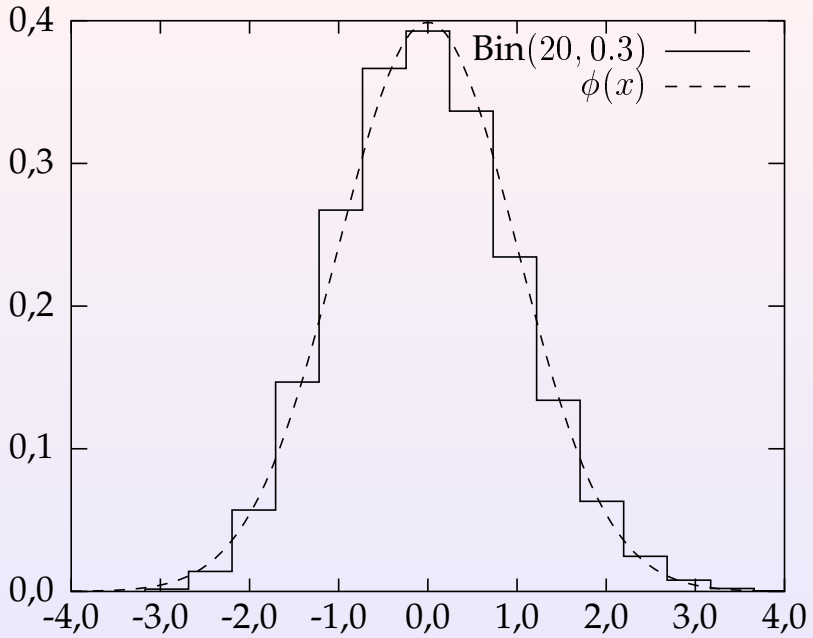
Normalverteilung als Grenzwert der Binomialverteilung

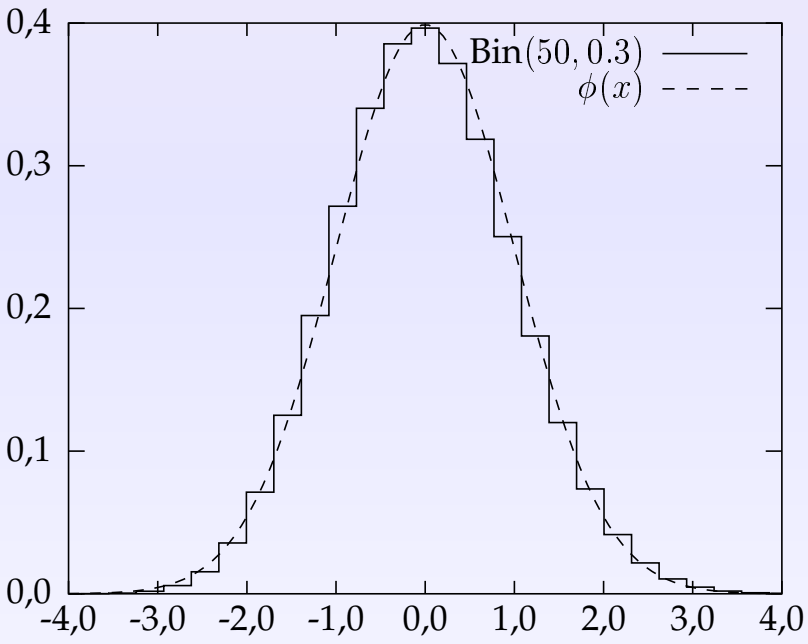
Korollar 46 ermöglicht, die Normalverteilung als Grenzwert der Binomialverteilung aufzufassen. Die folgende Aussage ist eine Konsequenz von Korollar 46:

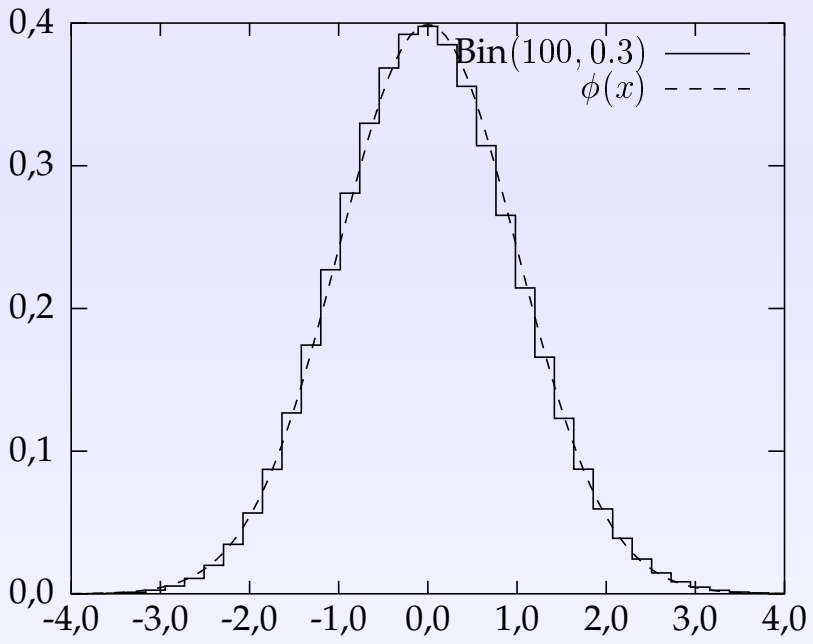
Korollar 47: Sei $H_n \sim \mathrm{Bin}(n,p)$ eine binomialverteilte Zufallsvariable. Die Verteilung von H_n/n konvergiert gegen $\mathcal{N}(p, p(1-p)/n)$ für $n \longrightarrow \infty$.





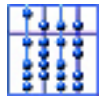


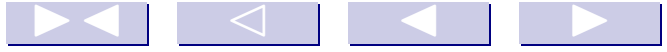




Vergleich von Binomial- und Normalverteilung

Historisch gesehen entstand Korollar 46 vor Satz 45. Für den Fall $p = 1/2$ wurde Korollar 46 bereits von Abraham DeMoivre (1667–1754) bewiesen. DeMoivre war gebürtiger Franzose, musste jedoch aufgrund seines protestantischen Glaubens nach England fliehen. Dort wurde er unter anderem Mitglied der Royal Society, erhielt jedoch niemals eine eigene Professur. Die allgemeine Formulierung von Korollar 46 geht auf Pierre Simon Laplace (1749–1827) zurück. Allerdings vermutet man, dass die Lösung des allgemeinen Falls $p \neq 1/2$ bereits DeMoivre bekannt war.





Elementarer Beweis des Grenzwertsatzes von DeMoivre für $p = 1/2$

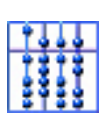
Wir betrachten die Wahrscheinlichkeit $\Pr[a \leq H^*_{2n} \leq b]$ für $p = 1/2$ und $a, b \in \mathbb{R}$ mit $a \leq b$. Wenn die Verteilung von H^*_{2n}, wie in Korollar 46 angegeben, gegen $\mathcal{N}(0,1)$ konvergiert, so sollte $\Pr[a \leq H^*_{2n} \leq b] \approx \int_a^b \varphi(t)\,\mathrm{d}\,t$ für genügend große n gelten.

Wir schreiben $f(n) \sim g(n)$ für $\lim_{n\to\infty} f(n)/g(n) = 1$, wollen also zeigen:

$$\Pr[a \leq H^*_{2n} \leq b] \sim \int_a^b \varphi(t)\,\mathrm{d}\,t.$$

Da für $H_{2n} \sim \mathrm{Bin}(2n, 1/2)$ gilt, dass $\mathbb{E}[H_{2n}] = n$ und $\mathrm{Var}[H_{2n}] = n/2$ ist, erhalten wir

$$H^*_{2n} = \frac{H_{2n} - n}{\sqrt{n/2}},$$



und es folgt

$$\begin{aligned}\Pr[a \leq H^*_{2n} \leq b] &= \Pr[n + a\sqrt{n/2} \leq H_{2n} \leq n + b\sqrt{n/2}] \\ &= \sum_{i \in I_n} \Pr[H_{2n} = n + i]\end{aligned}$$

für $I_n := \{z \in \mathbb{Z} \mid a\sqrt{n/2} \leq z \leq b\sqrt{n/2}\}$.

Damit ist

$$\Pr[a \leq H^*_{2n} \leq b] = \sum_{i \in I_n} \underbrace{\binom{2n}{n+i} \cdot \left(\frac{1}{2}\right)^{2n}}_{=:p_{n,i}}.$$

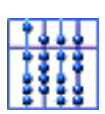
Es gilt

$$p_n^* := \max_i p_{n,i} = \binom{2n}{n} \cdot \left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{(n!)^2} \cdot \left(\frac{1}{2}\right)^{2n},$$

und mit der Stirling'schen Approximation für $n!$

$$p_n^* \sim \frac{(2n)^{2n} \cdot e^{-2n} \cdot \sqrt{2\pi \cdot 2n}}{(n^n \cdot e^{-n} \cdot \sqrt{2\pi n})^2} \cdot \left(\frac{1}{2}\right)^{2n} = \frac{1}{\sqrt{\pi n}}.$$

Approximieren wir nun die $p_{n,i}$ durch p_n^* so entsteht dabei ein Fehler, den wir mit $q_{n,i} := \frac{p_{n,i}}{p_n^*}$ bezeichnen.

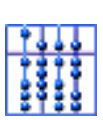


Für $i > 0$ gilt

$$q_{n,i} = \frac{\binom{2n}{n+i} \cdot \left(\frac{1}{2}\right)^{2n}}{\binom{2n}{n} \cdot \left(\frac{1}{2}\right)^{2n}} = \frac{(2n)! \cdot n! \cdot n!}{(n+i)! \cdot (n-i)! \cdot (2n)!}$$

$$= \frac{\prod_{j=0}^{i-1}(n-j)}{\prod_{j=1}^{i}(n+j)} = \prod_{j=1}^{i} \frac{n-j+1}{n+j} = \prod_{j=1}^{i} \left(1 - \frac{2j-1}{n+j}\right).$$

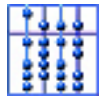
Wegen der Symmetrie der Binomialkoeffizienten gilt $q_{n,-i} = q_{n,i}$, womit auch der Fall $i < 0$ abgehandelt ist.



Man macht sich leicht klar, dass $1 - 1/x \leq \ln x \leq x - 1$ für $x > 0$ gilt. Damit schließen wir, dass

$$\begin{aligned}
\ln\left(\prod_{j=1}^{i}\left(1-\frac{2j-1}{n+j}\right)\right) &= \sum_{j=1}^{i}\ln\left(1-\frac{2j-1}{n+j}\right) \\
&\leq -\sum_{j=1}^{i}\frac{2j-1}{n+j} \leq -\sum_{j=1}^{i}\frac{2j-1}{n+i} \\
&= -\frac{i(i+1)-i}{n+i} = -\frac{i^2}{n}+\frac{i^3}{n(n+i)} \\
&= -\frac{i^2}{n}+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

da $i = \mathcal{O}(\sqrt{n})$ für $i \in I_n$.



Ebenso erhalten wir

$$\begin{aligned}
\ln\left(\prod_{j=1}^{i}\left(1-\frac{2j-1}{n+j}\right)\right) &\geq \sum_{j=1}^{i}\left(1-\left(1-\frac{2j-1}{n+j}\right)^{-1}\right)\\
&= \sum_{j=1}^{i}\frac{-2j+1}{n-j+1} \geq -\sum_{j=1}^{i}\frac{2j-1}{n-i}\\
&= -\frac{i^2}{n-i} = -\frac{i^2}{n}-\mathcal{O}\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Wegen $e^{\pm\mathcal{O}(1/\sqrt{n})} = 1 \pm o(1)$ folgt daraus

$$q_{n,i} \sim e^{-i^2/n}.$$

Damit schätzen wir nun $\Pr[a \leq H^*_{2n} \leq b]$ weiter ab:

$$\Pr[a \leq H^*_{2n} \leq b] \;=\; \sum_{i \in I_n} p^*_n \cdot q_{n,i} \;\sim\; \underbrace{\frac{1}{\sqrt{\pi n}} \cdot \sum_{i \in I_n} e^{-i^2/n}}_{=:S_n}.$$

Mit $\delta := \sqrt{2/n}$ können wir die Summe S_n umschreiben zu

$$S_n = \frac{1}{\sqrt{2\pi}} \cdot \sum_{i \in I_n} \delta e^{-(i\delta)^2 \cdot \frac{1}{2}}.$$

Diese Summe entspricht einer Näherung für $\int_a^b \varphi(t)\,\mathrm{d}\,t = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\,\mathrm{d}\,t$ durch Aufteilung der integrierten Fläche in Balken der Breite δ. Für $n \longrightarrow \infty$ konvergiert die Fläche der Balken gegen das Integral, d. h. $S_n \sim \int_a^b \varphi(t)\,\mathrm{d}\,t$. *q. e. d.*